Arabic Word Sense Disambiguation Using Wikipedia
Marwah Alian, Arafat Awajan and Akram Al-Kouz
Pages 61 – 66
DOI: http://dx.doi.org/10.21700/ijcis.2016.108

Arabic Word Sense Disambiguation Using Wikipedia

Marwah Alian, Arafat Awajan, Akram Al-Kouz
Department of Computer Science, Princess Sumaya University for Technology
Amman, Jordan
Marwah2001@yahoo.com, [Awajan,Akram]@ psut.edu.jo

Abstract: *In this research we introduce a new approach for Arabic word disambiguation by utilizing Wikipedia as the lexical resource for disambiguation. The nearest context for an ambiguous word is selected using Vector Space Model and cosine similarity between the word's context and the retrieved senses from Wikipedia. Three experiments have been conducted to evaluate the proposed approach, two experiments use the first retrieved sentence for each sense from Wikipedia but they use different Vector Space Model while the third experiment use the first paragraph for the retrieved sense from Wikipedia. The experiments show that using the first retrieved paragraph is better than the first retrieved sentence and the use of Tf-Idf VSM is better than using raw frequency VSM.*

Keywords: *Arabic Word Disambiguation; Disambiguation Resource; Vector Space Model; Arabic WordNet; Arabic Wkikpedia.*

Received: July 30, 2016 | ***Revised:*** August 10, 2016 | ***Accepted:*** August 25, 2016

1. Introduction

One of the most difficult problems in Natural Language Processing is the capability to identify what a word means with respect to its context. The technique that is used to find the appropriate sense of a word with ambiguous meaning considering its context is called Word Sense Disambiguation (WSD). [3][11]. It is ubiquitous across all languages but it has greater challenges in Semitic languages like Arabic language. WSD is considered as an AI-complete problem [2] and it is required in several applications such as machine translation [8][10], Information retrieval [1][7] and information extraction [5].

A word may denote different meanings in two different sentences because of ambiguity feature in human languages. For example, in English language the word bass may have different meaning according to the context in which it comes into view such as:

I can hear bass sounds.

They like grilled bass.

The word bass in the first sentence means musical instruments, while in the second sentence means type of a fish [11]. Furthermore, world knowledge is required in order to identify that the sense of 'bass' is a fish and not a musical instrument in the second sentence. This is because one would grill a fish not an instrument.

The methods used to disambiguate sense of word may be classified into three categories according to the techniques used: Knowledge-based, Supervised and Unsupervised methods where each category has a number of methods that are used.

Knowledge-based approaches use dictionaries while the supervised approaches are based on hand labeled data that are typically lexical samples. The unsupervised approaches use the information found in un-annotated corpora to distinguish the word meaning. [18]

In order to measure the similarity between two texts, each text is represented mathematically using a Vector Space Model (VSM). If the text that contains a sense of an ambiguous word and the query text have similar column vector, then the two compared texts have a similarity in their meaning. [13].

Many researches have been done for WSD in English but there are few researches for Arabic Word Sense Disambiguation (AWSD). Therefore, this work is looking for developing a new approach for AWSD using the knowledge-based approach, where the text is preprocessed and the ambiguous words senses are retrieved from Wikipedia then the retrieved senses and the tested text are represented as vectors where the cosine for the angle between the two vectors is computed. The cosine similarity measure is used in order to select the most appropriate meaning for an ambiguous word.

This paper is organized as follows: Section 2 gives an overview of the related work as well as a brief history of what have been proposed during the last few years

in AWSD. The proposed approach is discussed in section 3. Section 4 covers the evaluation experiments and the results are presented. Finally, the conclusion is presented in section 5.

2. Related Work

Since 1940's several researches have been introduced to solve the ambiguity of words in many languages [18]. However, the number of researches in Arabic word sense disambiguation (AWSD) is limited and the first approach was introduced in 2002 by [6] where they propose an unsupervised mechanism which utilizes the observation that words which have identical translation usually have similar dimension of meaning. Their algorithm considers a correct sense is strengthening by the semantic similarity of other words that share identical dimension of meaning. Here we will present the most recent work in AWSD.

One of the knowledge-based approaches is used in [14] where an evaluation for the variants of the Lesk algorithm is conducted to disambiguate Arabic words. They use the dictionary and perform the original Lesk algorithm then they experiment the modifications that they made to the lesk algorithm. They use similarity measures to identify how two concepts in Arabic Wordnet are relatedly similar. The modified Lesk algorithm gives a precision of 67%.

While in [15] they introduce a new approach in order to solve AWSD problem using Genetic Algorithm (GA), they call it GAWSD. They test their approach using a sample text in Arabic then they make a comparison with naïve Bayes classifier. The results show that their approach gives better performance than naïve classifier.

Other techniques are used in AWSD like hybrid techniques and fuzzy logic, for example, in [16] a hybrid technique is proposed for AWSD. The hybrid technique integrates unsupervised and knowledge-based methods in order to give a correct word meaning. By comparison, the hybrid technique outperforms other techniques in terms of precision. However, in [17] a fuzzy logic based technique is proposed to build a new classifier in order to be utilized in AWSD. In this work, two fuzzy logic classifiers are constructed and compared with a Naïve Byes classifier. The classifiers are used to identify the most possible senses for a word that has ambiguity. They identify a list of ambiguous words that consist of ten ambiguous words; they collect them from other researches handling the same problem. They argue that fuzzy logic considers the overlapping through different senses and that fuzzy logic deals with ambiguity and vagueness. Their experimental results show that fuzzy classifier gives more accurate results.

Recently, in [20] A new method is proposed for AWSD based on the global and local context of an ambiguous word where the correct sense is considered as the one that has a closer semantic similarity to both local and global context where local context is specified by the neighborhood of an ambiguous word, and the global context is specified by the whole text. Their proposed approach provides an accuracy of 74%.

Although in [21] they utilize both English WordNet and Arabic WordNet depending on machine translation for terms and they select the closest concept for an ambiguous word using the relationships between the ambiguous word and the different concepts in local context. In their experiments they use different machine learning and feature selection techniques to evaluate their method. The results of their system show that the proposed approach outperforms other techniques for AWSD.

Previous researches use different data sets with limited size and they are not available which makes it difficult to evaluate their results. For this reason, we propose a AWSD technique that considers open source resources for disambiguation such as Wikipedia to find the appropriate meaning of a given word by comparing the vector space Of the ambiguous text with the vector space calculated for different pages where the word occurs in Wikipedia which provide different possible meanings.

3. The Proposed Method

The proposed approach is based on comparing different possible meaning of the word in Wikipedia with the actual text then deciding the appropriate sense using similarity measurement. It consists of three phases. The first phase preprocesses the text and determines ambiguous words according to Arabic WordNet. The second phase searches Wikipedia for these ambiguous words and extracts the texts related to different meanings. The last phase compares VSM of the actual text with those extracted texts from Wikipedia and use a similarity measure to define the final meaning of the word. Figure 1 illustrates the processes in the proposed approach for all phases.

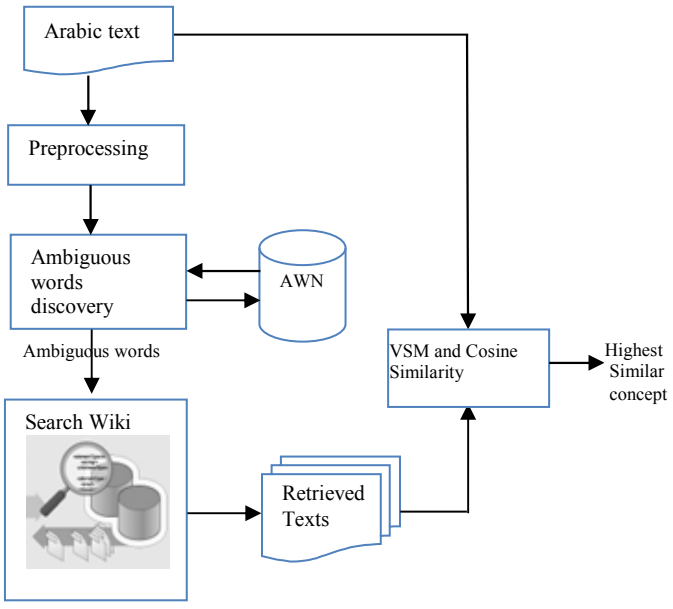

Figure 1: The Framework of the proposed approach

In the first phase, preprocessing is applied to the input text where tokenization and removing prefixes and suffixes are performed, then each token (word) is tested by Arabic WordNet (AWN) in order to obtain ambiguous words; according to AWN a word is ambiguous if it has more than one sense and it is non-ambiguous if it has only one sense.

The ambiguous words that were determined in phase one are passed to this phase to retrieve their senses from Arabic Wikipedia as a resource for disambiguation. The retrieved senses are concepts used to get summary text about these concepts through Wikipedia. Then these retrieved texts are passed to phase three.

In the last phase, each text in the retrieved texts from Wikipedia is represented as vectors and it is compared to the word's context, which is the input text to the system, using cosine similarity. The concept that has the most cosine similarity is considered as the appropriate sense for the ambiguous word.
In the following subsections more details about Arabic WordNet, Arabic Wikipedia, Vector space model and similarity measurement are given.

3.1 Arabic WordNet (AWN)

Arabic WordNet (AWN) is constructed for modern standard Arabic as lexical resource and it is based in its construction on the Princeton WordNet. It is organized into 11,269 synsets where a word can belong to one or more synsets. The number of words count in AWN is 23,481 words. [9][20]

3.2 Arabic Wikipedia

Arabic Wikipedia is the Arabic version of Wikipedia which is a cooperative Web encyclopedia that consists of pages. A Wikipage provides information about a specific sense or name entity [12]. Arabic Wikipedia contains over 400,000 articles and it is the first Semitic language that surpasses 100,000 articles [23]. However, the size of Arabic Wikipedia is considered relatively small if we take into consideration the number of Arabic speakers globally. [19]

3.3 Vector Space Model

Vector space model is a mathematical representation for documents. It is used in the measurement of documents similarity.

The steps of mathematical processing and construction of VSM are summarized in [4]; starting by frequency calculation then transforming raw frequency, reducing dimensionality and finally computing similarity.

Two hypotheses are considered in this research; Statistical semantics hypothesis and Bag of words hypothesis. The statistical hypothesis considers that two texts are considered similar if they have similar vectors in text frequency matrix [22] while the Bag of words hypothesis is based on words frequencies in a text. If the text and tested text have similar column vector, then this is an indication that they have similarity in their meaning [13].

Counter Vector is based on a mathematical concept called bag; which is a set where duplicates exist [13].
For example, {f; f; g; h; h; h} is called a bag which can be presented by a counter vector Y= [2; 1; 3] where 2 means that f is counted twice in the bag.
In order to measure the similarity between two frequency vectors is to take the cosine of the angle between them. The cosine similarity between two vectors x and y is computed as in Equation (1) [13].

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \ \|y\|} \qquad (1)$$

Where x.y is the inner product of the vectors, and

$$\|x\| = \text{sqrt}((x[0])^2 + (x[1])^2 \ldots).$$

If the result equals one then the two vectors are similar and if it is zero then there is no similarity.

4. Experiments and Results

For evaluating the proposed approach, we conduct three different experiments. In the first experiment, we limit the text extracted to one sentence and compare the raw frequency VSM of the text containing the ambiguous word with one sentence of Wiki text containing the word.
In the second experiment, we use also one retrieved sentence from Wikipedia but the retrieved texts and the actual text are represented using a Tf-Idf vector space model.

In the third experiment, the extracted text from Wikipedia is expanded to be one paragraph containing the ambiguous word and Tf-Idf vector space model is applied. Table 1 shows the senses given by both Wikipedia and AWN for each tested ambiguous word in the experiments.

Table 1: Word senses from AWN and Wikipedia

Ambiguous word	Word Senses in AWN	Word Senses in Wikipedia
عين	ينبوع ، عميل سري، جاسوس، رقيب	عين الإنسان ، حرف العين، عين الحسد، مخيم عين بيت الماء ، عين زبيدة ، عين جالوت، معجم العين ، عين النسر : بلدة ، عين النسر: فيلم، عين الشرقية: قرية ببانياس، عين الشرقية: بلدة سورية، راس العين :جماعة قروية
معلم	معلم، مهذب، مربي، صفة ، طابع ، معلم، مظهر	المدرس ، المزار او المعلم السياحي، المعلم التذكاري، دار المعلمين، جامع حي المعلمين، معهد ضبا ط، المعلمية، شخصية خيالية المعلم بالروسية، بندر معلم بالفارسية: قرية، المعالم الجغرافية، نظام معلومات الاسماء الجغرافية
رب	إله ، معبود، صاحب، مالك	الرباني بالعبرية: سيد ، الرب: الاله، الحرب، بيت الربية ، جلسة التشهد، وربة: جزيرة، عبد ربه ، الربان، أحمد عبدالرب، أم الرب: قرية
سيارة	سيارة، مركبة	تصنيف السيارات ، السيارة: المركبة، صناعة السيارات ، ، السيارة carsالسيارات: فيلم مفخخة، دفع رباعي، لوحة مركبات ، قائمة الطرقات السيارة في تونس، سباق سيارات، سيارة مدينة
رئاسة	قيادة، رئاسة ، زعامة	الرئيس، النظام الرئاسي، قوات الحماية الرئاسية، الرئاس: قرية، الوسام الرئاسي للحرية، اللواء الأول حماية رئاسية، الانتخابات الرئاسية: تونس، الانتخابات الرئاسية: سوريا، وزير شؤون الرئاسة، دار الرئاسة
فصل	عزل، تجزئة، تقطيع، تقسيم ، فصل كتاب، نمرة مسرحية، كلام معاد، روتين، موسم	فصول السنة، الفصل التعليمي، العزل العنصري، فصل السلطات، فصل (منطق)، عملية الفصل في الهندسة الكيميائية،الفصل الكهربائي، الفصل(قرية)، الفصل (حبيش)، قمع الفصل، ميثاق الامم المتحدة: فصل، فصل النظائر، الصيف، الشتاء
طالب	أقام دعوى، اشترط، احتاج، استدعى، تلميذ، متعلم، دارس، باحث	الطالبية: قرية – الجيزة، الطالبية : قرية في كفر الزيات، الطالب: المتعلم، بلدية الطالب، أبو طالب، لكمة بن سعيد طالب: قرية، أبو طالب: قرية، علي بن أبي طالب، الطالب العربي: دائرة في الجزائرة، طالب بن أبي طالب، بني طالب: قرية

Table 2 shows the ambiguous word, the tested text and the highest similar sense that is retrieved from Arabic Wikipedia in the three experiments.

Table 2: Experiment results

Ambiguous word	Tested text	Highest similar sense Exp1	Highest similar sense Exp2	Highest similar sense Exp3
عيـــن	في البستان عين تجري	عين الشرقية: بلدة سورية	عين العفريت	عين العفريت
	تستقبلنا مدينة راس العين ببهاء يتختلف عن بقية المدن	عين الحسد	عين العفريت	عين العفريت
	أرسلت عينا على العدو	عين الحسد	عين العفريت	عين العفريت
	أصابه الناس بالعين فجلس مريضاً في بيته	عين النسر : بلدة	عين العفريت	عين العفريت
	نظر إليه نظرة فأصابه	عين الحسد	عين الحسد	عين الحسد
	عين الولد تؤلمه	عين الإنسان	عين العفريت	عين العفريت
رب	للبيت رب يحميه	بيت الربية	بيت الربية	بيت الربية
	المسافر يدعو رب العالمين ليسلمه	الرباني بالعبرية: سيد	عبد ربه	عبد ربه
	للمسلم رب يدعوه ليدبر له أمره	الرباني بالعبرية: سيد	الرب: الاله	الرب: الاله
رئاسة	تجري الانتخابات لرئاسة الجمهورية في منتصف الشهر	الانتخابات الرئاسية: سوريا	الانتخابات الرئاسية: سوريا	الانتخابات الرئاسية: تونس
	تكون رئاسة الفريق للشخص الأقوى	دار الرئاسة	دار الرئاسة	الانتخابات الرئاسية: سوريا
معلم	كانت ضربة معلم	المدرس	المعلمية	المعالم الجغرافية
	يوصف بأنه معلم في مهنة الحدادة	المدرس	دار المعلمين	دار المعلمين
	شرح معلم العلوم الدرس	المدرس	المدرس	المدرس
فصل	ترتفع الحرارة في هذا الفصل	ميثاق الامم المتحدة: فصل	قمع الفصل	قمع الفصل
	أنهى أحمد متطلبات هذا الفصل	الفصل (حبيش)	الفصل (حبيش)	الفصل التعليمي
	لابد من فصل الجزيئات في هذه التجربة	الصيف	الصيف	الصيف
طالب	طالب الحق يلح في السؤال	الطالب: المتعلم	الطالب: المتعلم	أبو طالب: قرية
	على الطالب ان ينهي عددا من المواد الدراسية	الطالب: المتعلم	الطالب: المتعلم	الطالب:المتعلم
سيارة	ان الكواكب السيارة تدور في فلكها	سيارة مدينة	سيارة مدينة	دفع رباعي
	استخدم السيارة للذهاب الى عملك	سيارة مدينة	سيارة مدينة	السيارة: المركبة

The experiments show that retrieving the first paragraph from Wikipedia is better than retrieving the first sentence. Also, applying VSM with Tf-Idf is better than using raw frequency VSM.

We can explain the unsatisfied results of wrong sense to a word by problems in Arabic WordNet and Arabic Wikipedia which. The limited coverage in AWN and the size of documents in Arabic Wikipedia as well as noise in these documents lead to errors in the assigned sense to the ambiguous word. Also, the proposed approach concentrates on the global context represented by the vector space model while the local context of the word is not taken into consideration.

5. Conclusion

The problems of Arabic WordNet (AWN) such as noise, precision and limited coverage compared to English WordNet leads to enrich with another resource for disambiguation as Wikipedia. Therefore, in this research a new approach for Arabic word disambiguation is proposed utilizing Wikipedia and applying Vector Space Model as a mathematical representation for documents.

The proposed method is implemented and evaluated using cosine similarity measure. Three experiments are performed; the first one use one retrieved sentence for the senses from Wikipedia and gives the correct sense for words (عين، رئاسة، طالب، معلم، سيار) in one tested text and two matches for the word (رب), while in the second experiment, Tf-Idf vector space model is used and the results matches the correct sense for the words (عين، رئاسة، رب، طالب ، معلم، سيارة) The first and second experiment did not match the correct sense for the word (فصل) but it is matched for one tested text in the third experiment which retrieved a paragraph for each sense from Wikipedia. Also, the third experiment gives the correct sense for one tested text with the words (عين، رئاسة ،طالب، معلم، رب، سيارة).

As a future work the local context of the ambiguous word will be considered with the word's global context to produce better results.

References

[1] Schütze, H., and Pedersen, J. "Information Retrieval Based on Word Senses", *in Proceedings of Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, 1995, pp. 161-175.

[2] MALLERY, J. C.. Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers. *Ph.D. dissertation. MIT Political Science Department, Cambridge, MA*. 1988

[3] Ide, N., Véronis, J., Word Sense Disambiguation: The State of the Art, *Computational Linguistics, Vol. 24, No. 1*, 1998, pp. 1-40.

[4] Lowe, W. Towards a theory of semantic space. *In Proceedings of the Twenty-_rst Annual Conference of the Cognitive Science Society*, 2001, pp. 576-581.

[5] Jacquemin, B., Brun, C., and Boux, C. ,Enriching a Text by Semantic Disambiguation for Information Extraction, *in Proceedings of the Workshop on Using Semantics for Information Retrieval and Filtering in the 3rd International Conference in Language Resources and Evaluation (LREC)*, 2002

[6] M. Diab, P. Resnik, An unsupervised method for word sense tagging using parallel corpora. *in Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02),* Philadelphia,2002, pp. 255–262.

[7] Stokoe, C., Oakes, M., and Tait, J., Word Sense Disambiguation in Information Retrieval Revisited, *in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 159-166.

[8] Carpaut, M., and Wu, D., Word Sense Disambiguation vs. Statistical Machine Translation, *in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005, pp. 387-394.

[9] Elkateb, S., Black W., Vossen P., Farwell D., Rodríguez H., Pease A., Alkhalifa M., Arabic WordNet and the Challenges of Arabic, *In proceedings of Arabic NLP/MT Conference, London, U.K*, 2006.

[10] Chan, Y., Ng, H., and Chiang, D., Word Sense Disambiguation Improves Statistical Machine Translation*, in Proceedings of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL),* 2007, pp. 33-40.

[11] Roberto Navigli, Word Sense Disambiguation: A Survey, *ACM Computing Surveys, Vol. 41, No. 2, ACM Press*, pp:1-69, 2009.

[12] Simone Paolo Ponzetto, Roberto Navigli, Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1522–1531.

[13] Peter D. Turney,Patrick Pantel, From Frequency to Meaning:Vector Space Models of Semantics, *Journal of Articial Intelligence Research 37* , 2010, pp. 141-188.

[14] A. Zouaghi, L. Merhbene, M. Zrigui, Word Sense disambiguation for Arabic language using the variants of the Lesk algorithm, *in the proceedings of WORLDCOMP'11*, 2011, pp. 561-567.

[15] Mohamed El Bachir Menai, Wojdan Alsaeedan, Genetic algorithm for Arabic word sense disambiguation, *13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE*, 2012, pp. 195-200

[16] Anis Zouaghi, A Hybrid Approach for Arabic Word Sense Disambiguation, *International Journal of Computer Processing Of Languages, Vol. 24, No. 2* , 2012 , pp.133–151

[17] Madeeh Nayer El-Gedawy, 2013, Using Fuzzifiers to Solve Word Sense Ambiguation in Arabic Language, *International Journal of Computer Applications, Volume 79 – No2,* 2013.

[18] Alok Ranjan Pal, Diganta Saha, 2015, Word Sense Disambiguation: A Survey, *International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3*, July 2015

[19] Ahmad, Abdullah, Arabic Wikipedia: Why it lags behind, *Asfar e-Journal (London, UK)*. 2013.

[20] Nadia Bouhriz, Faouzia Benabbou, El Habib Ben Lahmar, Word Sense Disambiguation Approach for Arabic Text,*(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 4*, 2016, pp.381-385.

[21] Meryeme Hadni, Said El Alaoui, Abdelmonaime Lachkar, Word Sense Disambiguation for Arabic Text Categorization, *The International Arab Journal of Information Technology, Vol. 13, No. 1A*, 2016, pp.215-222.

[22] Weaver, W., Translation. In Locke, W., & Booth, D. (Eds.), Machine Translation of Languages: Fourteen Essays. *MIT Press, Cambridge, MA*, 1955.

[23] Arabic Wikipedia definition retrieved at 22 June 2016from : https://en.wikipedia.org/wiki/Arabic_Wikipedia

Marwah Alian received her M.S.c degree in Computer Science in 2007 from Jordan University. Since 2015, she is working towards her PhD at the Department of Computer Science of Princess Sumaya University for Technology. She has a number of publications in e-learning systems, data mining, and mobile applications.

Arafat Awajan is a full professor of computer science at Princess Sumaya University for Technology (PSUT). He is currently the dean of the King Hussein faculty of Computing Sciences. He received his PhD degree in computer science from the University of Franche, Comte, France in 1987. He held different academic positions at the Royal Scientific Society and Princess Sumaya University for Technology. He was appointed as the chair of the Computer Science Department (2000-2003) and the chair of the Computer Graphics and Animation Department (2005-2006) at PSUT. He had been the dean of the King Hussein School for Information Technology from 2004 to 2007, the Dean of Student Affairs from 2011-2014 and the director of the Information Technology Center in the Royal Scientific Society from 2008-2010. His research interests include natural language processing, text compression, and image processing.

Akram Alkouz is assistant professor at the department of Computer Science/Princess Sumaya University for Technology since September 2013. Het Got His PhD from Berlin Institute of Technology (TU-Berlin). His Research Interests include Arabic NLP, Big Data and Machine Learning.